Chapter 1

Summarizing Data

When determining how to appropriately analyze any collection of data, the first consideration must be the characteristics of the data themselves. Little is gained by employing analysis procedures which assume that the data possess characteristics which in fact they do not. The result of such false assumptions may be that the interpretations provided by the analysis are incorrect, or unnecessarily inconclusive. Therefore we begin this book with a discussion of the common characteristics of water resources data. These characteristics will determine the selection of appropriate data analysis procedures.

One of the most frequent tasks when analyzing data is to describe and summarize those data in forms which convey their important characteristics. "What is the sulfate concentration one might expect in rainfall at this location"? "How variable is hydraulic conductivity"? "What is the 100 year flood" (the 99th percentile of annual flood maxima)? Estimation of these and similar summary statistics are basic to understanding data. Characteristics often described include: a measure of the center of the data, a measure of spread or variability, a measure of the symmetry of the data distribution, and perhaps estimates of extremes such as some large or small percentile. This chapter discusses methods for summarizing or describing data.

This first chapter also quickly demonstrates one of the major themes of the book -- the use of robust and resistant techniques. The reasons why one might prefer to use a resistant measure, such as the median, over a more classical measure such as the mean, are explained.

The data about which a statement or summary is to be made are called the **population**, or sometimes the **target population**. These might be concentrations in all waters of an aquifer or stream reach, or all streamflows over some time at a particular site. Rarely are all such data available to the scientist. It may be physically impossible to collect all data of interest (all the water in a stream over the study period), or it may just be financially impossible to collect them. Instead, a subset of the data called the **sample** is selected and measured in such a way that conclusions about the sample may be extended to the entire population. Statistics computed from the sample are only inferences or estimates about characteristics of the population, such as location, spread, and skewness. Measures of location are usually the sample mean and sample median. Measures of spread include the sample standard deviation and sample interquartile range. Use of the term "sample" before each statistic explicitly demonstrates that these only estimate the population value, the population mean or median, etc. As sample estimates are far more common than measures based on the entire population, the term "mean" should be interpreted as the "sample mean", and similarly for other statistics used in this book. When population values are discussed they will be explicitly stated as such.

1.1 Characteristics of Water Resources Data

Data analyzed by the water resources scientist often have the following characteristics:

1. A lower bound of zero. No negative values are possible.
2. Presence of 'outliers', observations considerably higher or lower than most of the data, which infrequently but regularly occur. outliers on the high side are more common in water resources.
3. Positive skewness, due to items 1 and 2. An example of a skewed distribution, the lognormal distribution, is presented in figure 1.1. Values of an observation on the horizontal axis are plotted against the frequency with which that value occurs. These density functions are like histograms of large data sets whose bars become infinitely narrow. Skewness can be expected when outlying values occur in only one direction.
4. Non-normal distribution of data, due to items 1 - 3 above. Figure 1.2 shows an important symmetric distribution, the normal. While many statistical tests assume data follow a normal distribution as in figure 1.2, water resources data often look more like figure 1.1. In addition, symmetry does not guarantee normality. Symmetric data with more observations at both extremes (heavy tails) than occurs for a normal distribution are also non-normal.
5. Data reported only as below or above some threshold (censored data). Examples include concentrations below one or more detection limits, annual flood stages known only to be lower than a level which would have caused a public record of the flood, and hydraulic heads known only to be above the land surface (artesian wells on old maps).
6. Seasonal patterns. Values tend to be higher or lower in certain seasons of the year.

7. Autocorrelation. Consecutive observations tend to be strongly correlated with each other. For the most common kind of autocorrelation in water resources (positive autocorrelation), high values tend to follow high values and low values tend to follow low values.
8. Dependence on other uncontrolled variables. Values strongly covary with water discharge, hydraulic conductivity, sediment grain size, or some other variable.

Methods for analysis of water resources data, whether the simple summarization methods such as those in this chapter, or the more complex procedures of later chapters, should recognize these common characteristics.

1.2 Measures of Location

The mean and median are the two most commonly-used measures of location, though they are not the only measures available. What are the properties of these two measures, and when should one be employed over the other?

1.2.1 Classical Measure -- the Mean

The mean $(\overline{X})$ is computed as the sum of all data values X_i, divided by the sample size n:

$$\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n} \qquad [1.1]$$

For data which are in one of k groups, equation [1.1] can be rewritten to show that the overall mean depends on the mean for each group, weighted by the number of observations n_i in each group:

$$\overline{X} = \sum_{i=1}^{n} \overline{X}_i \frac{n_i}{n} \qquad [1.2]$$

where $\overline{X}_i$ is the mean for group i. The influence of any one observation X_j on the mean can be seen by placing all but that one observation in one "group", or

$$\overline{X} = \overline{X}_{(j)} \frac{(n-1)}{n} + X_j \bullet \frac{1}{n}.$$

$$= \overline{X}_{(j)} + (X_{(j)} - \overline{X}_{(j)}) \bullet \frac{1}{n}. \qquad [1.3]$$

where $\overline{X}_{(j)}$ is the mean of all observations excluding X_j. Each observation's influence on the overall mean $\overline{X}$ is $(X_j - \overline{X}_{(j)})$, the distance between the observation and the mean excluding that observation. Thus all observations do not have the same influence on the mean. An 'outlier' observation, either high or low, has a much greater influence on the overall mean $\overline{X}$ than does a more 'typical' observation, one closer to its $\overline{X}_{(j)}$.

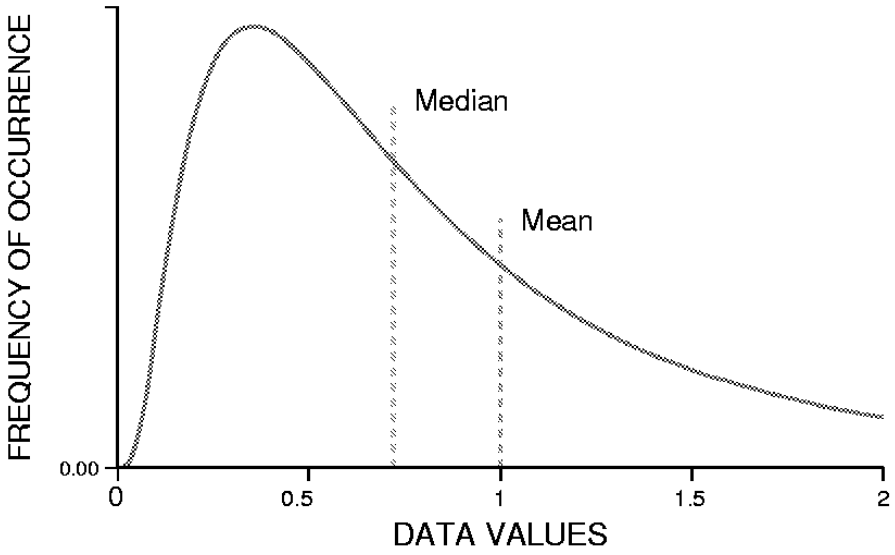

Figure 1.1 Density Function for a Lognormal Distribution

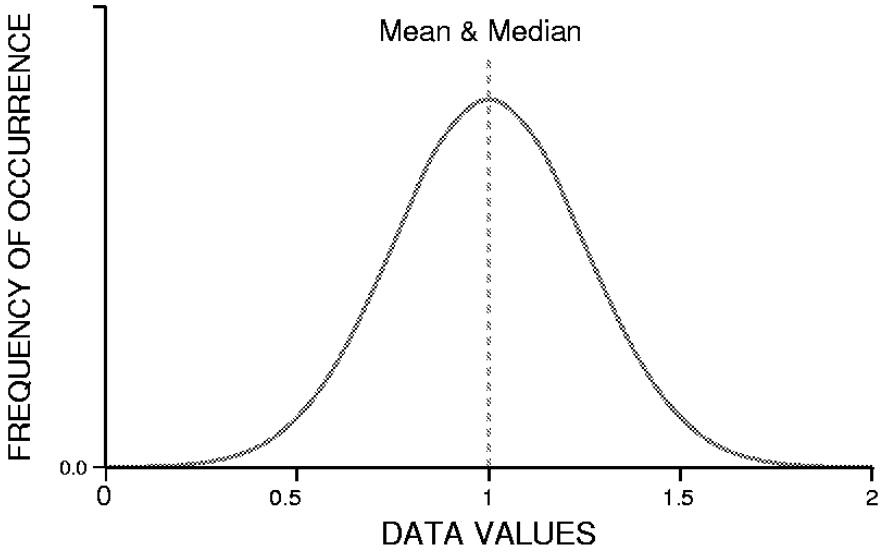

Figure 1.2 Density Function for a Normal Distribution

Another way of illustrating this influence is to realize that the mean is the balance point of the data, when each point is stacked on a number line (figure 1.3a). Data points further from the center exert a stronger downward force than those closer to the center. If one point near the center were removed, the balance point would only need a small adjustment to keep the data set in balance. But if one outlying value were removed, the balance point would shift dramatically (figure 1.3b). This sensitivity to the magnitudes of a small number of points in the data set defines why the mean is not a "resistant" measure of location. It is not resistant to changes in the presence of, or to changes in the magnitudes of, a few outlying observations.

When this strong influence of a few observations is desirable, the mean is an appropriate measure of center. This usually occurs when computing units of mass, such as the average concentration of sediment from several samples in a cross-section. Suppose that sediment concentrations closer to the river banks were much higher than those in the center. Waters represented by a bottle of high concentration would exert more influence (due to greater mass of sediment per volume) on the final concentration than waters of low or average concentration. This is entirely appropriate, as the same would occur if the stream itself were somehow mechanically mixed throughout its cross section.

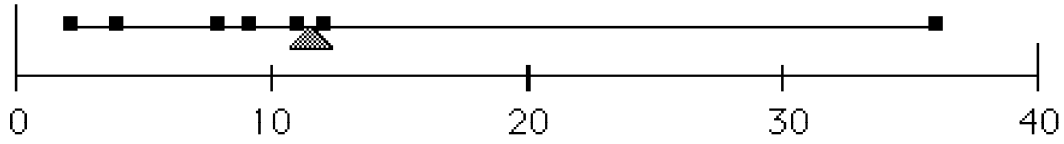

Figure 1.3a The mean (triangle) as balance point of a data set.

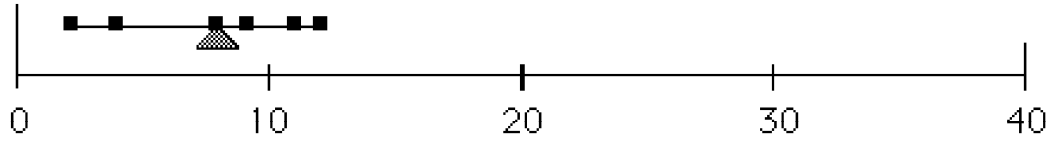

Figure 1.3b Shift of the mean downward after removal of outlier.

1.2.2 Resistant Measure -- the Median

The median, or 50th percentile $P_{0.50}$, is the central value of the distribution when the data are ranked in order of magnitude. For an odd number of observations, the median is the data point which has an equal number of observations both above and below it. For an even number of observations, it is the average of the two central observations. To compute the median, first

rank the observations from smallest to largest, so that x_1 is the smallest observation, up to x_n, the largest observation. Then

$$\text{median } (P_{0.50}) = X_{(n+1)/2} \qquad \text{when n is odd, and}$$

$$\text{median } (P_{0.50}) = \frac{1}{2}\left(X_{(n/2)} + X_{(n/2)+1}\right) \qquad \text{when n is even.} \qquad [1.4]$$

The median is only minimally affected by the magnitude of a single observation, being determined solely by the relative order of observations. This resistance to the effect of a change in value or presence of outlying observations is often a desirable property. To demonstrate the resistance of the median, suppose the last value of the following data set (a) of 7 observations were multiplied by 10 to obtain data set (b):

Example 1:

(a)	2 4 8 9 11 11 12	$\overline{X} = 8.1$	$P_{.50} = 9$
(b)	2 4 8 9 11 11 120	$\overline{X} = 23.6$	$P_{.50} = 9$

The mean increases from 8.1 to 23.6. The median, the $\frac{(7+1)}{2}$ th or 4th lowest data point, is unaffected by the change.

When a summary value is desired that is not strongly influenced by a few extreme observations, the median is preferable to the mean. One such example is the chemical concentration one might expect to find over many streams in a given region. Using the median, one stream with unusually high concentration has no greater effect on the estimate than one with low concentration. The mean concentration may be pulled towards the outlier, and be higher than concentrations found in most of the streams. Not so for the median.

1.2.3 Other Measures of Location

Three other measures of location are less frequently used: the mode, the geometric mean, and the trimmed mean. The mode is the most frequently observed value. It is the value having the highest bar in a histogram. It is far more applicable for grouped data, data which are recorded only as falling into a finite number of categories, than for continuous data. It is very easy to obtain, but a poor measure of location for continuous data, as its value often depends on the arbitrary grouping of those data.

The geometric mean (GM) is often reported for positively skewed data sets. It is the mean of the logarithms, transformed back to their original units.

$$GM = \exp(\overline{Y}), \qquad \text{where } Y_i = \ln(X_i) \qquad [1.5]$$

(in this book the natural, base e logarithm will be abbreviated **ln**, and its inverse e^x abbreviated **exp(x)**). For positively skewed data the geometric mean is usually quite close to the median. In fact, when the logarithms of the data are symmetric, the geometric mean is an unbiased estimate

of the median. This is because the median and mean logarithms are equal, as in figure 1.2. When transformed back to original units, the geometric mean continues to be an estimate for the median, but is not an estimate for the mean (figure 1.1).

Compromises between the median and mean are available by trimming off several of the lowest and highest observations, and calculating the mean of what is left. Such estimates of location are not influenced by the most extreme (and perhaps anomalous) ends of the sample, as is the mean. Yet they allow the magnitudes of most of the values to affect the estimate, unlike the median. These estimators are called "trimmed means", and any desirable percentage of the data may be trimmed away. The most common trimming is to remove 25 percent of the data on each end -- the resulting mean of the central 50 percent of data is commonly called the "trimmed mean", but is more precisely the 25 percent trimmed mean. A "0% trimmed mean" is the sample mean itself, while trimming all but 1 or 2 central values produces the median. Percentages of trimming should be explicitly stated when used. The trimmed mean is a resistant estimator of location, as it is not strongly influenced by outliers, and works well for a wide variety of distributional shapes (normal, lognormal, etc.). It may be considered a weighted mean, where data beyond the cutoff 'window' are given a weight of 0, and those within the window a weight of 1.0 (see figure 1.4).

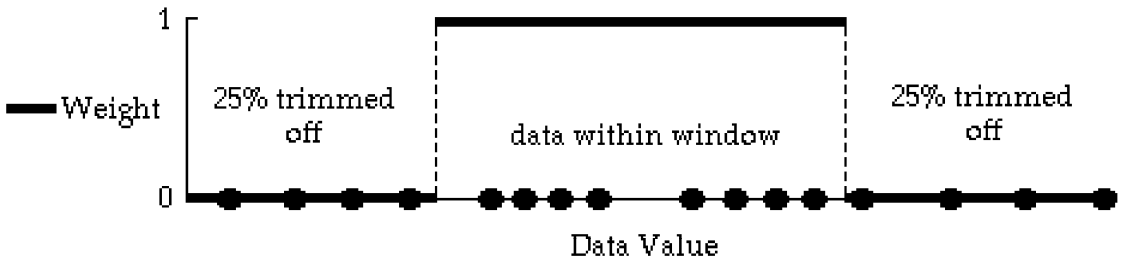

Figure 1.4. Window diagram for the trimmed mean

1.3 Measures of Spread

It is just as important to know how variable the data are as it is to know their general center or location. Variability is quantified by measures of spread.

1.3.1 Classical Measures

The sample variance, and its square root the sample standard deviation, are the classical measures of spread. Like the mean, they are strongly influenced by outlying values.

$$s^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{(n-1)} \qquad \text{sample variance} \qquad [1.6]$$

$$s = \sqrt{s^2} \qquad \text{sample standard deviation} \qquad [1.7]$$

They are computed using the squares of deviations of data from the mean, so that outliers influence their magnitudes even more so than for the mean. When outliers are present these measures are unstable and inflated. They may give the impression of much greater spread than is indicated by the majority of the data set.

1.3.2 Resistant Measures

The interquartile range (IQR) is the most commonly-used resistant measure of spread. It measures the range of the central 50 percent of the data, and is not influenced at all by the 25 percent on either end. It is therefore the width of the non-zero weight window for the trimmed mean of figure 1.4.

The IQR is defined as the 75th percentile minus the 25th percentile. The 75th, 50th (median) and 25th percentiles split the data into four equal-sized quarters. The 75th percentile ($P_{.75}$), also called the upper quartile, is a value which exceeds no more than 75 percent of the data and is exceeded by no more than 25 percent of the data. The 25th percentile ($P_{.25}$) or lower quartile is a value which exceeds no more than 25 percent of the data and is exceeded by no more than 75 percent. Consider a data set ordered from smallest to largest: X_i, i =1,...n. Percentiles (P_j) are computed using equation [1.8]

$$P_j = X_{(n+1)\bullet j} \qquad [1.8]$$

where n is the sample size of X_i, and
j is the fraction of data less than or equal to the percentile value (for the 25th, 50th and 75th percentiles, j= .25, .50, and .75).

Non-integer values of (n+1)•j imply linear interpolation between adjacent values of X. For the example 1 data set given earlier, n=7, and therefore the 25th percentile is $X_{(7+1)\bullet .25}$ or X_2 = 4, the second lowest observation. The 75th percentile is X_6 , the 6th lowest observation, or 11. The IQR is therefore 11−4 = 7.

One resistant estimator of spread other than the IQR is the Median Absolute Deviation, or MAD. The MAD is computed by first listing the absolute value of all differences |d| between each observation and the median. The median of these absolute values is then the MAD.

$$\text{MAD}\ (X_i) = \text{median}\ |d_i|, \qquad \text{where } d_i = X_i - \text{median}\ (X_i) \qquad [1.9]$$

Comparison of each estimate of spread for the Example 1 data set is as follows. When the last value is changed from 12 to 120, the standard deviation increases from 3.8 to 42.7. The IQR and the MAD remain exactly the same.

data	2	4	8	9	11	11	12	IQR = 11 – 4 = 7
$(X_i - \overline{X})^2$	37.2	16.8	0.01	0.81	8.41	8.41	15.2	$s^2 = (3.8)^2$
$\lvert d_i = X_i - P_{.50} \rvert$	7	5	1	0	2	2	3	MAD=median$\lvert d_i \rvert$=2
data	2	4	8	9	11	11	120	IQR = 11 – 4 = 7
$(X_i - \overline{X})^2$	37.2	16.8	0.01	0.81	8.41	8.41	12,522	$s^2 = (42.7)^2$
$\lvert d_i = X_i - P_{.50} \rvert$	7	5	1	0	2	2	111	MAD=median$\lvert d_i \rvert$=2

1.4 Measures of Skewness

Hydrologic data are typically skewed, meaning that data sets are not symmetric around the mean or median, with extreme values extending out longer in one direction. The density function for a lognormal distribution shown previously as figure 1.1 illustrates this skewness. When extreme values extend the right tail of the distribution, as they do with figure 1.1, the data are said to be skewed to the right, or positively skewed. Left skewness, when the tail extends to the left, is called negative skew.

When data are skewed the mean is not expected to equal the median, but is pulled toward the tail of the distribution. Thus for positive skewness the mean exceeds more than 50 percent of the data, as in figure 1.1. The standard deviation is also inflated by data in the tail. Therefore, tables of summary statistics which include only the mean and standard deviation or variance are of questionable value for water resources data, as those data often have positive skewness. The mean and standard deviation reported may not describe the majority of the data very well. Both will be inflated by outlying observations. Summary tables which include the median and other percentiles have far greater applicability to skewed data. Skewed data also call into question the applicability of hypothesis tests which are based on assumptions that the data have a normal distribution. These tests, called parametric tests, may be of questionable value when applied to water resources data, as the data are often neither normal nor even symmetric. Later chapters will discuss this in much detail, and suggest several solutions.

1.4.1 Classical Measure of Skewness

The coefficient of skewness (g) is the skewness measure used most often. It is the adjusted third moment divided by the cube of the standard deviation:

$$g = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \frac{(x_i - \overline{X})^3}{s^3} \qquad [1.10]$$

A right-skewed distribution has positive g; a left-skewed distribution has negative g. Again, the influence of a few outliers is important -- an otherwise symmetric distribution having one outlier will produce a large (and possibly misleading) measure of skewness. For the example 1 data, the g skewness coefficient increases from –0.5 to 2.6 when the last data point is changed from 12 to 120.

1.4.2 Resistant Measure of Skewness

A more resistant measure of skewness is the quartile skew coefficient qs (Kenney and Keeping, 1954):

$$qs = \frac{(P_{.75} - P_{.50}) - (P_{.50} - P_{.25})}{P_{.75} - P_{.25}} \qquad [1.11]$$

the difference in distances of the upper and lower quartiles from the median, divided by the IQR. A right-skewed distribution again has positive qs; a left-skewed distribution has negative qs. Similar to the trimmed mean and IQR, qs uses the central 50 percent of the data. For the example 1 data, qs = (11–9) – (9–4) / (11–4) = –0.43 both before and after alteration of the last data point. Note that this resistance may be a liability if sensitivity to a few observations is important.

1.5 Other Resistant Measures

Other percentiles may be used to produce a series of resistant measures of location, spread and skewness. For example, the 10 percent trimmed mean can be coupled with the range between the 10th and 90th percentiles as a measure of spread, and a corresponding measure of skewness:

$$qs_{.10} = \frac{(P_{.90} - P_{.50}) - (P_{.50} - P_{.10})}{P_{.90} - P_{.10}} \qquad [1.12]$$

to produce a consistent series of resistant statistics. Geologists have used the 16th and 84th percentiles for many years to compute a similar series of robust measures of the distributions of sediment particles (Inman, 1952). However, measures based on quartiles have become generally standard, and other measures should be clearly defined prior to their use. The median, IQR, and quartile skew can be easily summarized graphically using a boxplot (see Chapter 2) and are familiar to most data analysts.

1.6 Outliers

Outliers, observations whose values are quite different than others in the data set, often cause concern or alarm. They should not. They are often dealt with by throwing them away prior to describing data, or prior to some of the hypothesis test procedures of later chapters. Again, they should not. Outliers may be the most important points in the data set, and should be investigated further.

It is said that data on the Antarctic ozone "hole", an area of unusually low ozone concentrations, had been collected for approximately 10 years prior to its actual discovery. However, the automatic data checking routines during data processing included instructions on deleting "outliers". The definition of outliers was based on ozone concentrations found at mid-latitudes. Thus all of this unusual data was never seen or studied for some time. If outliers are deleted, the risk is taken of seeing only what is expected to be seen.

Outliers can have one of three causes:

1. a measurement or recording error.
2. an observation from a population not similar to that of most of the data, such as a flood caused by a dam break rather than by precipitation.
3. a rare event from a single population that is quite skewed.

The graphical methods of the Chapter 2 are very helpful in identifying outliers. Whenever outliers occur, first verify that no copying, decimal point, or other obvious error has been made. If not, it may not be possible to determine if the point is a valid one. The effort put into verification, such as re-running the sample in the laboratory, will depend on the benefit gained versus the cost of verification. Past events may not be able to be duplicated. If no error can be detected and corrected, **outliers should not be discarded based solely on the fact that they appear unusual**. Outliers are often discarded in order to make the data nicely fit a pre-conceived theoretical distribution such as the normal. There is no reason to suppose that they should! The entire data set may arise from a skewed distribution, and taking logarithms or some other transformation may produce quite symmetrical data. Even if no transformation achieves symmetry, outliers need not be discarded. Rather than eliminating actual (and possibly very important) data in order to use analysis procedures requiring symmetry or normality, procedures which are resistant to outliers should instead be employed. If computing a mean appears of little value because of an outlier, the median has been shown to be a more appropriate measure of location for skewed data. If performing a t-test (described later) appears invalidated because of the non-normality of the data set, use a rank-sum test instead.

In short, let the data guide which analysis procedures are employed, rather than altering the data in order to use some procedure having requirements too restrictive for the situation at hand.

1.7 Transformations

Transformations are used for three purposes:

1. to make data more symmetric,
2. to make data more linear, and
3. to make data more constant in variance.

Some water resources scientists fear that by transforming data, results are derived which fit preconceived ideas. Therefore, transformations are methods to 'see what you want to see' about the data. But in reality, serious problems can occur when procedures assuming symmetry, linearity, or homoscedasticity (constant variance) are used on data which do not possess these required characteristics. Transformations can produce these characteristics, and thus the use of transformed variables meets an objective. Employment of a transformation is not merely an arbitrary choice.

One unit of measurement is no more valid a priori than any other. For example, the negative logarithm of hydrogen ion concentration, pH, is as valid a measurement system as hydrogen ion concentration itself. Transformations like the square root of depth to water at a well, or cube root of precipitation volume, should bear no more stigma than does pH. These measurement scales may be more appropriate for data analysis than are the original units. Hoaglin (1988) has written an excellent article on hidden transformations, consistently taken for granted, which are in common use by everyone. Octaves in music are a logarithmic transform of frequency. Each time a piano is played a logarithmic transform is employed! Similarly, the Richter scale for earthquakes, miles per gallon for gasoline consumption, f-stops for camera exposures, etc. all employ transformations. In the science of data analysis, the decision of which measurement scale to use should be determined by the data, not by preconceived criteria. The objectives for use of transformations are those of symmetry, linearity and homoscedasticity. In addition, the use of many resistant techniques such as percentiles and nonparametric test procedures (to be discussed later) are invariant to measurement scale. The results of a rank-sum test, the nonparametric equivalent of a t-test, will be exactly the same whether the original units or logarithms of those units are employed.

1.7.1 The Ladder of Powers

In order to make an asymmetric distribution become more symmetric, the data can be transformed or re-expressed into new units. These new units alter the distances between observations on a line plot. The effect is to either expand or contract the distances to extreme observations on one side of the median, making it look more like the other side. The most commonly-used transformation in water resources is the logarithm. Logs of water discharge, hydraulic conductivity, or concentration are often taken before statistical analyses are performed.

Transformations usually involve power functions of the form $y = x^{\theta}$, where x is the untransformed data, y the transformed data, and θ the power exponent. In figure 1.5 the values of θ are listed in the "ladder of powers" (Velleman and Hoaglin, 1981), a useful structure for determining a proper value of θ.

As can be seen from the ladder of powers, any transformations with θ less than 1 may be used to make right-skewed data more symmetric. Constructing a boxplot or Q-Q plot (see Chapter 2) of the transformed data will indicate whether the transformation was appropriate. Should a logarithmic transformation overcompensate for right skewness and produce a slightly left-skewed distribution, a 'milder' transformation with θ closer to 1, such as a square-root or cube-root transformation, should be employed instead. Transformations with $\theta > 1$ will aid in making left-skewed data more symmetric.

Figure 1.5
"LADDER OF POWERS"
(modified from Velleman and Hoaglin, 1981)

Use	θ	Transformation	Name	Comment
		•		higher powers can be used
		•		
for (−)		•		
skewness	3	x^3	cube	
	2	x^2	square	
	1	x	original units	no transformation
	1/2	$\sqrt{x}$	square root	commonly used
	1/3	$\sqrt[3]{x}$	cube root	commonly used
	0	log(x)	logarithm	commonly used. Holds the place of x^0
for (+) skewness				
	−1/2	$-1/\sqrt{x}$	reciprocal root	the minus sign preserves order of observations
	−1	−1/x	reciprocal	
	−2	$-1/x^2$		
		•		
		•		
		•		lower powers can be used

However, the tendency to search for the 'best' transformation should be avoided. For example, when dealing with several similar data sets, it is probably better to find one transformation which works reasonably well for all, rather than using slightly different ones for each. It must be remembered that each data set is a sample from a larger population, and another sample from the same population will likely indicate a slightly different 'best' transformation. Determination of 'best' in great precision is an approach that is rarely worth the effort.

Exercises

1.1 Yields in wells penetrating rock units without fractures were measured by Wright (1985), and are given below. Calculate the

a) mean
b) trimmed mean
c) geometric mean
d) median
e) compare these estimates of location. Why do they differ?

Unit well yields (in gal/min/ft) in Virginia (Wright, 1985)

0.001	0.030	0.10	0.003	0.040	0.454
0.007	0.041	0.49	0.020	0.077	1.02

1.2 For the well yield data of exercise 1.1, calculate the

a) standard deviation
b) interquartile range
c) MAD
d) skew and quartile skew.

Discuss the differences between a through c.

1.3 Ammonia plus organic nitrogen (in mg/L) was measured in samples of precipitation by Oltmann and Shulters (1989). Some of their data are presented below. Compute summary statistics for these data. Which observation might be considered an outlier? How should this value affect the choice of summary statistics used

a) to compute the mass of nitrogen falling per square mile.
b) to compute a "typical" concentration and variability for these data?

0.3	0.9	0.36	0.92	0.5	1.0
0.7	9.7	0.7	1.3		